A SIMPLE SPATIALLY WEIGHTED MEASURE OF TEMPORAL STABILITY FOR DATA WITH LIMITED TEMPORAL OBSERVATIONS

J. Piburn [a,*], R. Stewart [a], A. Morton [a]

[a] Oak Ridge National Laboratory, 1 Bethel Valley Road Oak Ridge, TN 37830 - (piburnjo, stewartrn, mortonam)@ornl.gov

KEY WORDS: Time Series, Approximate Entropy, Spatio-Temporal, Exploratory Spatial Data Analysis (ESDA), Data Mining.

ABSTRACT:

Identifying erratic or unstable time-series is an area of interest to many fields. Recently, there have been successful developments towards this goal. These new developed methodologies however come from domains where it is typical to have several thousand or more temporal observations. This creates a challenge when attempting to apply these methodologies to time-series with much fewer temporal observations such as for socio-cultural understanding, a domain where a typical time series of interest might only consist of 20-30 annual observations. Most existing methodologies simply cannot say anything interesting with so few data points, yet researchers are still tasked to work within in the confines of the data. Recently a method for characterizing instability in a time series with limited temporal observations was published. This method, Attribute Stability Index (ASI), uses an approximate entropy based method to characterize a time series' instability. In this paper we propose an explicitly spatially weighted extension of the Attribute Stability Index. By including a mechanism to account for spatial autocorrelation, this work represents a novel approach for the characterization of space-time instability. As a case study we explore national youth male unemployment across the world from 1991-2014.

1. INTRODUCTION

In socio-cultural research, identifying instability in spatio-temporal trends is critical for understanding global dynamics and finding areas of potential concern or intervention. A common constraint however, is the limited temporal observations that data of interest are typical of having. This leaves many of the advances from the time series data mining literature unable to contribute in a meaningful capacity to socio-cultural research, as they rely on having hundreds, if not thousands, of temporal observations for responsible application. Although potentially useful in other fields, this paper represents an attempt to develop a time series data mining technique specifically designed with the application constraints and operational definitions of identifying spatio-temporal instability in socio-cultural research.

1.1 Two Dimensions of Stability

For operational definitions in socio-cultural research, instability is marked by two characteristics, 1) how widely varying the values are and 2) how predictable that variance is from one observation to the next. How widely varying the values of a time series are can be considered simply through variance. However, moment statistics such as variance do not consider the order of the observations in their summaries. For example, a vector that alternates regularly between the values of 5 and 10 has the same variance as a vector that takes the value of 5 or 10 each with a probability of ½. Variance is thus a necessary, but not sufficient, measure of instability. To address how predictable changes in a time series are, various instantiations of entropy have seen much success. Most of the entropic methods were born out of applications with numerous temporal observations, such as finance or heart rate monitoring, and thus were not designed with observational constraints in mind. A notable example otherwise is approximate entropy (ApEn). Introduced by Pincus (1991), ApEn is a computational approximation of Kolmogorov-Sinai entropy and is used to measure the amount of regularity and unpredictability in a time series. Due to its approximate nature ApEn is not burdened by the need for numerous observations and thus is particularly well suited for time series' consisting of limited data points. ApEn however is no panacea. While useful in measuring how predictable changes are in a time series, ApEn does not consider how widely varying the changes are, just the regularity of a change. Furthermore, ApEn is a parametric model that puts requirements on the user to know at what value to set the input parameters. Of particular responsiveness is the value of r which sets the threshold of what the model considers a change. Anything above r is counted towards the measure and anything below it passes by as if nothing changed at all. This of course can widely alter the resulting ApEn value of time series if all changes in the series were just below this value compared to just above it.

Recently, Piburn, Morton, and Stewart (2017) proposed a methodology, Attribute Stability Index (ASI), that incorporates both the magnitude and predictability of changes for a time series with limited observations. They show by approximating the integral of ApEn with respect to r you can produce stability curves and a summary value that allow for exploration and insights into the temporal stability of a time series.

In this paper, we extend this methodology to incorporate the temporal stability information from neighbouring locations for each area of interest. This extension allows you to find local anomalies in spatial-temporal stability. That is to say, each location's value now incorporates not only how unstable a location is through time, but also how unstable this temporal instability is through space. We provide a motivating example of this methodology by exploring national youth male unemployment across the world from 1991-2014.

2. METHODOLOGY

In this section, we will define what is needed to arrive at the spatial extension of the ASI. Section 2.1 describes ApEn before section 2.2 details the ASI and its spatial extension. A comprehensive understanding of Section 2.1 isn't needed to continue with section 2.2, but is included for the interested reader.

A note on notation used in this section. Unless a mistake that passed unseen, all scaler values are indicated by an italicized lower case letter, such as x, or for referential continuity purposes, a capital Greek letter, such as Φ. All vectors are denoted by a

*Corresponding author

bold lower case letter, such as **x** and any matrices are indicated by a bold capitalized letter, such as **W**.

2.1 Approximate Entropy

Given a time series of interest $\mathbf{x} = (x_i, x_{i+1}, \dots, x_n)$ construct a sequence of vectors, $\mathbf{s}_i$, $\mathbf{s}_{i+1}, \dots, \mathbf{s}_{n-m+1}$ where $\mathbf{s}_i = (x_i, x_{i+1}, \dots, x_{i+m-1})$. $\mathbf{s}_i$ is simply a subset of the original time series of interest **x**, starting at the i^{th} observation and continuing forward until the defined ending, the $i^{th} + m - 1$ observation. So for example, $\mathbf{s}_{i+1} = (x_{i+1}, \dots, x_{i+m-1})$. For each **s** vector we can now calculate the correlation dimension as follows

$$c_i^m(r) = (n - m + 1)^{-1} \sum_{j \neq i} \begin{cases} d_{ij} \leq r = 1 \\ d_{ij} > r = 0 \end{cases} \quad (1)$$

Where d_{ij} is defined as the following

$$d_{ij} = d(\mathbf{s}_i, \mathbf{s}_j) = \max_{k=1,2,\dots,m} (|x_{i+k-1} - x_{j+k-1}|) \quad (2)$$

$c_i^m(r)$ counts the number of times, as a percentage, that d_{ij} met or exceeded *r*. At this point we have $n - m + 1$ values of $c_i^m(r)$ representing our original times series **x**. Taking the sum of their logs and then normalizing we get a measure of the average $c_i^m(r)$ value for **x** given m and r. In Equation 3, Φ^m itself can be thought of as a measure of the fractal dimensionality of **x** at the scale of m. By deriving this value for **x** again but instead using $m + 1$, the approximate entropy estimate for **x** comes naturally out of the difference between them. This difference can be interpreted as how different is our measurement of **x** when we change the scale at which we measure **x**, similar in idea and formulation to a fractal dimension.

$$\Phi^m(r) = (n - m + 1)^{-1} \sum_{i=1}^{n-m+1} \log c_i^m(r) \quad (3)$$

$$\Theta(m, r) = \Phi^m(r) - \Phi^{m+1}(r) \quad (4)$$

For further details on approximate entropy please see Pincus (1995) and Richman and Moorman (2000).

Going forward we will set $m = 2$, as we are concerned with immediate change from one observation to the next, and drop it for notational simplicity. Do note however, that all remaining calculations hold at any value of m.

2.2 Attribute Stability Index

As mentioned above, even setting m to a fixed value still leaves the ApEn estimate for **x** dependent upon what level of r is chosen. If you think of ApEn as a function of r, you could plot its behaviour over all possible r values and see how the ApEn values for **x** respond. Furthermore, since the ApEn of any time series at $r = 0$ or $r > \max(d_{ij})$ will itself be 0, means that we can easily evaluate all possible none-zero ApEn outputs of this function. This is in fact exactly how the attribute stability index is calculated; it is the approximation of the definite integral of $\Theta(r)$ from $r = 0$ to the logical maximum value for each time series. By integrating over all values of r, we accomplish two things 1) we sidestep the problem of setting an arbitrary value of r and thus the sensitivity concern and 2) we can get a more complete picture of a time series' instability not only graphically but intuitively as well by incorporating all changes large and small. ApEn as defined in equation 4 is reintroduced in equation 8, but first a few preliminary calculations must be made.

Given the same time series we introduced in section 2.1, $\mathbf{x} = (x_i, x_{i+1}, x_{i+2}, \dots, x_n)$, the first step is to calculate the lagged difference of the vector with a lag of 1, this can be seen in equation 5.

$$\mathbf{x}_{lag} = \{(x_{i+1} - x_i), (x_{i+2} - x_{i+1}), \dots, (x_n - x_{n-1})\} \quad (5)$$

Once $\mathbf{x}_{lag}$ is defined, the absolute value of the maximum lag is calculated. This value is upper bound for the values of r used in the ApEn calculations, with the lower bound being 0 as stated above. The maximum lag is defined as the upper bound because any value of r that is greater than the largest lag by definition will result inan ApEn value of 0. Equation 6 uses this lag and an integer λ, which can set by the user as an input into the ASI calculations, to determine ρ, how large of a step to take between successive evaluations of ApEn. The larger the value of λ the closer the approximation will be to the definite integral. λ can be thought of as the resolution of the resulting approximation

$$\rho = \frac{\max|\mathbf{x}_{lag}|}{\lambda} \quad (6)$$

Using ρ and a vector of integers from 0 to λ we can construct a vector, **r** (Equation 7), that contains all of values of r for which we will evaluate ApEn used in the ASI calculation.

$$\begin{aligned} \mathbf{y} &= (0,1,2,3, \dots, \lambda) \\ \mathbf{r} &= \rho \mathbf{y} \end{aligned} \quad (7)$$

At this point the final ASI estimation can be calculated with your favourite numerical integration method, here we use the trapezoidal form (Equation 8).

$$\begin{aligned} \alpha &= \frac{1}{2} \sum_{k=1}^{N} (r_{k+1} - r_k) \cdot (\Theta(\mathbf{x}, r_{k+1}) + \Theta(\mathbf{x}, r_k)) \\ &\approx \int_0^{\rho\lambda} \Theta(\mathbf{x}) dr \end{aligned} \quad (8)$$

α then is the ASI scaler summary value for each time series of interest. If each of these time series are associated with some location in space, we can then extend our ASI estimates to incorporate this spatial information.

If we define $\boldsymbol{\alpha}$ as the vector of ASI values for all locations, we can calculate a measure of how stable a location is relative to its neighbours by taking the ratio κ_i of its ASI value to the average of its neighbour's values

$$\kappa_i = \frac{\alpha_i}{\boldsymbol{W_i \alpha}} \quad (9)$$

where the neighbours of area i and there contribution to the local neighbourhood mean ASI is given by a standard spatial weights matrix $\boldsymbol{W_i}$ (for example queen contiguity).

Th vector $\boldsymbol{\kappa}$ of κ_i provides a relative measure of how much more unstable a location is compared to its neighbours. A value above 1 indicates that a location has greater temporal instability then it's neighbourhood and a value of less than 1 indicating the opposite, a locations neighbourhood has greater temporal instability then the location itself. This however removes any magnitude differences between neighbourhoods, a location with an ASI of 2 and neighbourhood average of 1 would have the same value of a location with an ASI value of 20 and neighbourhood

average of 10. If this alone is an appropriate measure for the question you are trying to answer than **κ** could be used itself to explore the spatio-temporal relationships of the areas under study, however, for our particular problem we still want to consider the magnitude of the variability and with the ratio alone we do not get this information. To reintroduce the effects of the magnitude of the changes we see, we use **κ** to weight our original ASI vector **α**. Specifically, we scale each measure again according to Equation 10.

$$\alpha_{sp(i)} = \alpha_i \cdot \kappa_i \quad (10)$$

The vector $\boldsymbol{\alpha}_{sp}$ is our final vector of spatially weighted temporal stability measures $\alpha_{sp(i)}$. This approach has the desired effect of shrinking the ASI values of locations that although temporally unstable, they are more stable than their spatial neighbours and increasing the values of locations that while not alone largely unstable, locally they are much more unstable then their neighbours. At the extremes, where locations have both large (or both small) **α** and **κ,** the result is double down on their distinctiveness, increasing (or shrinking) even more the measure of their spatio-temporal stability.

3. CASE STUDY

As a case study of this methodology we explore the behaviour of youth male unemployment as a percent of male labour force ages 15-24 from 1991-2014 at the national level across the globe. For illustration purposes, the actual trends from five example countries are shown in Figure 1. These five countries have a range of different behaviours that provide insight into how ASI calculations behave. Botswana and France both have widely varying values with no clearly identifiable overall trend, Trinidad and Tobago has non-trivial changes in value from one observation to the next but with an overall downward trend, and finally Bahrain and China have values that tend to change in an irregular pattern but by smaller amounts. Since the goal of the ASI is to identify trends that are both irregular and widely varying, these examples will provide a better understanding of how individual trends are scored.

The non-spatially weighted ASI results can be investigated in two ways. First, we can look at the attribute stability curve, the behaviour of each locations ApEn values across values of r. Figure 2 shows these curves for the example countries in figure 1. The shape of the location' curve gives us insight into the nature of the instability in the trend. We can see that at very low values of r, China and Bahrain have higher ApEn values than France or Botswana, but then drop quickly to zero. This indicates that while the changes from year to year in China and Bahrain may be irregular, they do not vary by a large amount. France has the highest peak ApEn value of the example countries and maintains high ApEn values across a wider range of r values than that of the other countries except for Botswana. Although Botswana's peak ApEn value is not has high as that of France, the ApEn values stay higher over a much wider range of r values. The shape of Botswana's stability curve tells us that not only are changes from year to year irregular, the amount by which it changes is also irregular.

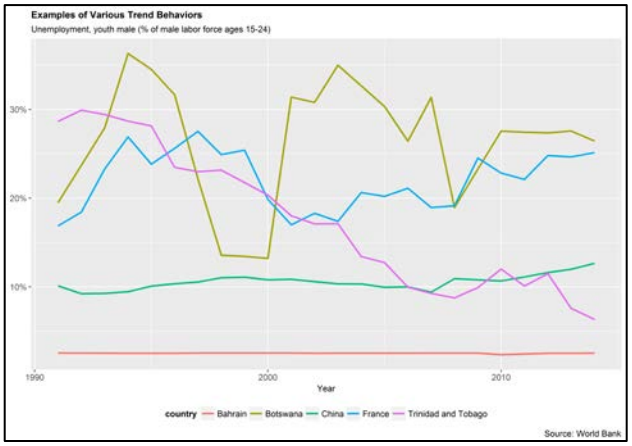

Figure 1. Examples of Various Trend Behaviours

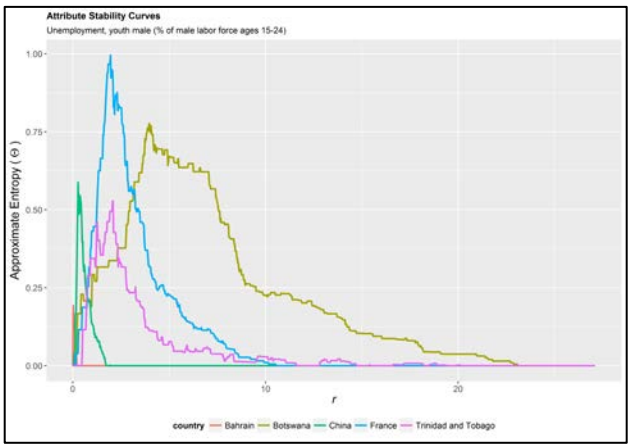

Figure 2. Attribute Stability Curves of Examples Trends

Expanding the results to all countries in the case study, the spatial distribution of the non-spatially weighted ASI scores can be seen in Figure 3.

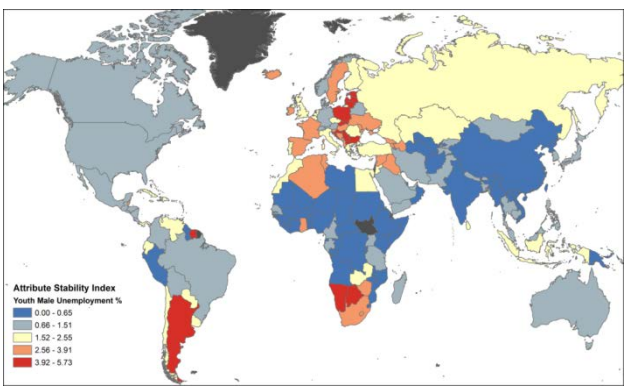

Figure 3. Attribute Stability Index

Spatial patterns immediately emerge including a large cluster of low ASI scores in central Africa and high scores in Eastern Europe. As for the example countries from above, Botswana falls into the highest break along with neighbouring Namibia, France is a member of the second highest grouping, while China is in the lowest category, in line with what we expected from inspecting the attribute stability curves. An important note is that the ASI values represent a temporal behaviour, not just a single value of the attribute in question. By summarising temporal behaviour into a single measure, it allows the spatial

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-IV-4-W2-47-2017 | © Authors 2017. CC BY 4.0 License.

distribution of temporal behaviours to be visualized on a static map, without the use of animation or multiple visualizations.

Now that we have considered the non-spatially weighted ASI values for each location we can start exploring the spatially weighted extension. For this example, we defined our spatial weights matrix with queen contiguity. Following this definition countries with 0 defined neighbours are excluded from the spatially weighted ASI study. Figure 4 compares each location's ASI, or α, value on the x axis to the average of its neighbours ASI on the y axis. The black diagonal line represents equal x and y values, that is locations where their ASI value is equal to the average of their neighbours. Locations above the line, $\boldsymbol{\kappa} > 1$, can be considered more stable through time then their average neighbour, while locations below the line, $\boldsymbol{\kappa} < 1$, are more unstable through time than their neighbourhood average.

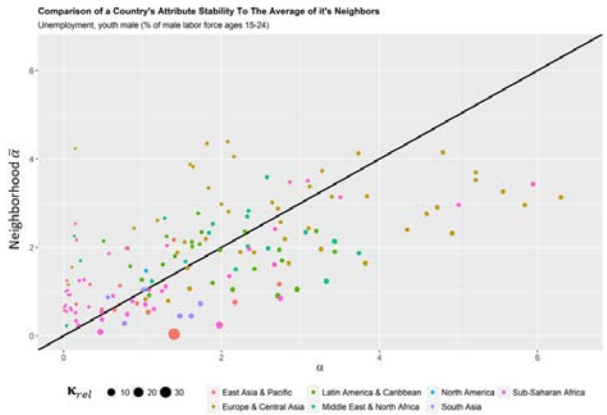

Figure 4. Comparison of a Country's ASI (α) to the Average of its Neighbours

Looking at the spatial distribution of our spatially weighted ASI values, $\boldsymbol{\alpha_{sp}}$, in Figure 5 a few distinct patterns emerge. The most prominent global pattern is that for the most part, the temporal stability of youth male unemployment in a country is relatively close in magnitude and predictability to its neighbours. The interesting stories of course are the countries that deviate from what is normal.

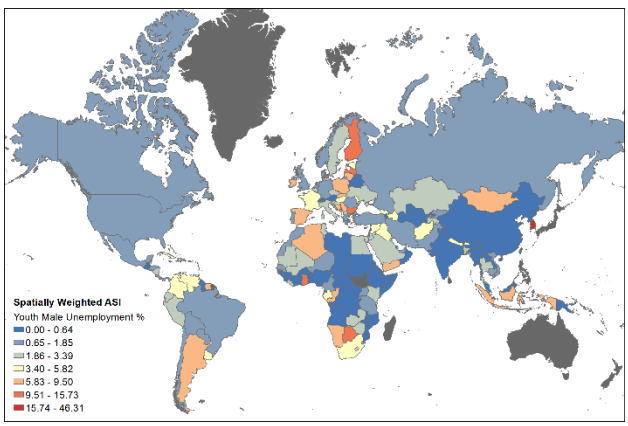

Figure 5. Spatially Weighted Attribute Stability Index ($\boldsymbol{\alpha_{sp}}$)

The country with the lowest, α_{sp}, the country that is the most stable in an unstable neighbourhood, is Belarus. Figure 6 shows the youth male unemployment trends for Belarus and all its neighbours.

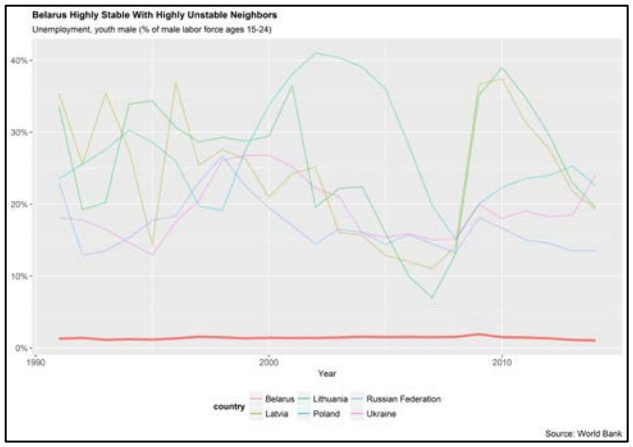

Figure 6. Belarus Identified with the Lowest α_{sp}

During the entire study period the rate of youth male unemployment in Belarus remained incredulously low and stable, ranging from a minimum of 1.06% in 2014 and a maximum of 1.9% in 2009. During the same period, the neighbour of Belarus saw multiple double digits swings from year to year.

On the opposite side, countries that are much more unstable through time than their neighbours, one country dwarfs the rest. South Korea's only neighbour (by our queen contiguity definition) is North Korea and North Korea's reported youth male unemployment trend is remarkably stable, bouncing between 12.5-12.7% from 1996-2014. Figure 7 shows these trends

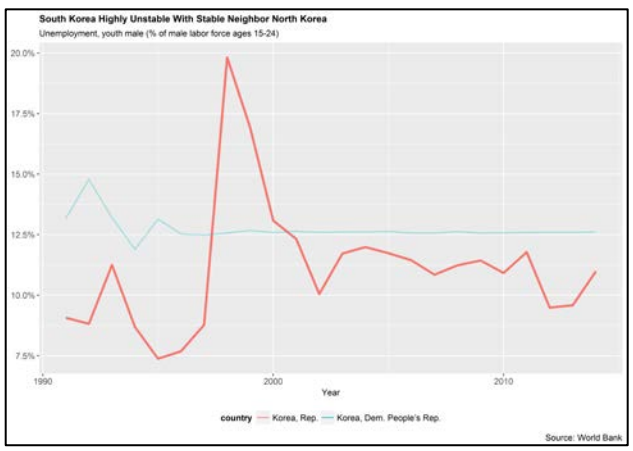

Figure 7. South Korea Identified with the Highest α_{sp}

An additional example of a country that is much more unstable over time than its neighbours is Ghana. The trend of Ghana and its neighbours are shown in Figure 8.

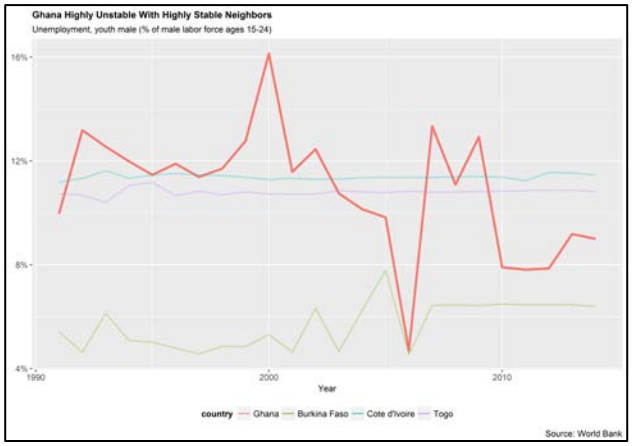

Figure 8. Ghana Identified with the 2nd Highest α_{sp}

4. CONCLUSION

In this paper, we proposed a spatially weighted extension to the attribute stability index, a method for investigating and quantifying instability in time series data with limited temporal observations; particularly as used in the field of spatio-temporal socio-cultural research where instability is understood to mean widely varying and irregular changes from one observation to the next. This methodology represents an additional tool for exploratory spatio-temporal data analysis and provides a novel technique for researchers to use in understanding spatio-temporal trends.

ACKNOWLEDGEMENTS

This manuscript has been authored by employees of UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the U.S. Department of Energy. Accordingly, the United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

REFERENCES

Piburn, J. O., Stewart, R.N, & Morton, A.M. (2017). An Approximate Entropy Based Approach for Quantifying Stability in Spatio-Temporal Data with Limited Temporal Observations. *21st International Conference on GeoComputation.* Leeds, U.K. *(accepted)*

Pincus, S. M. (1991). Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences, 88*(6), 2297-2301.

Pincus, S. M. (1995). Approximate entropy (ApEn) as a complexity measure. *Chaos: An Interdisciplinary Journal of Nonlinear Science, 5*(1), 110-117.

Richman, J. S., & Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology-Heart and Circulatory Physiology, 278*(6), H2039-H2049.